Review on Data Mining with Big Data

Savita Suryavanshi, Prof. Bharati Kale

***Abstract*— Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.**

***Index Terms*— Big Data, data mining, heterogeneity, autonomous sources, complex and evolving associations.**

I. INTRODUCTION

The data held by Australian Government agencies is both a national and government asset. It is also a potential source of opportunity. In this context, Australian Government agencies, like many other organizations, are aware of the challenges and opportunities that big data represents to the way they develop policy and deliver services to citizens. The purpose of this issues paper is to provide an opportunity to consider the range of opportunities presented to agencies in relation to the use of big data, and the emerging tools that allow us to better appreciate what it tells us, in the context of the potential concerns that this might raise. As an example, one of the major challenges facing agencies here is to leverage the value of big data sets while ensuring they continue to protect the privacy rights of the Australian public. The Australian Government is committed to protecting citizen's rights to privacy, and as part of that commitment, has recently strengthened the provisions of the Privacy Act. The Australian Government Information Management Office (AGIMO) acknowledges that big data, and its associated analytical tools, can provide a challenge to these rights, but believe that, with proper considerations, agencies will be able to use big data to develop better policies and deliver better services without compromising the privacy rights of the public. Our aim is to ensure that the use of the new technology and tools supporting big data will deliver benefits while maintaining compliance with privacy. To this end AGIMO will be working closely with the Office of the Australian Information Commissioner (OAIC), the Attorney General's Department (AGD) and experts across the public and private sectors as it develops a big data strategies

Savita Suryavanshi, Department of Computer Engineering, DPCOE, Wagholi,Pune, India.
Prof. Bharati Kale, Department of Computer Engineering, DPCOE, Wagholi,Pune, India.

The idea of "Big Data" is in the air. At the South by Southwest Interactive conference last month, it was probably the hot topic, dominating or surfacing in numerous panels, including one on which I spoke, on "Big Data: Privacy Threat or Business Model?"

Let's be clear on what we're talking about. The term refers to something more specific than the general fact that companies and government agencies are collecting lots of personal information about people. What it refers to is the fact that once you store up huge amounts of information, you can mine those databases to discover subtle patterns, correlations, or relationships that our brains can't perceive on their own because the scales involved are beyond our ability to process (either the time scales at work, or the sheer number of data points).Such data mining has been called "the macro scope"—like a telescope or microscope, making things visible to us that have never been visible before.

In many ways Big Data is just a new buzzword for data mining, which we and others have been grappling with since not long after 9/11. The New York Times, for example, wrote about it using the term "data mining" in this piece. A more recent (and much-discussed) article by Charles Duhigg in the New York Times offers a good example to keep in mind during discussions of the subject. The piece described how Target identifies customers who are pregnant (sometimes before their own family members know) by tracking customers' purchases and identifying patterns in their behavior. It then uses that insight to sell them baby-related goods.

II. PROBLEM DEFINITION

It incentivizes more collection of data and longer retention of it. If any and all data sets might turn out to prove useful for discovering some obscure but valuable correlation, you might as well collect it and hold on to it. In long run, the more useful big data proves to be, the stronger this incentivizing effect will be—but in the short run it almost doesn't matter; the current buzz over the idea is enough to do the trick.

III. PERSPECTIVE SOLUTION

Apache Hadoop! A Solution for Big Data!

Hadoop is an open source software framework that supports data-intensive distributed applications. Hadoop is licensed under the Apache v2 license. It is therefore generally known as Apache Hadoop. Hadoop has been developed, based on a paper originally written by Google on Map Reduce system and applies concepts of functional programming. Hadoop is written in the Java programming language and is the highest-level Apache project being constructed and used by a global community of contributors. Hadoop was developed by

Doug Cutting and Michael J. Cafarella. And just don't overlook the charming yellow elephant you see, which is basically named after Doug's son's toy elephant!

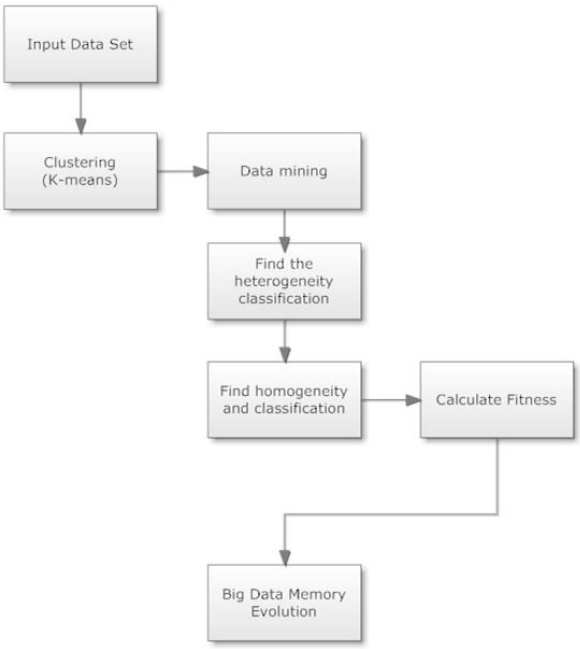
IV. RELATED WORK AND LITERATURE SURVEY

Dynamic networks have recently being recognized as a powerful abstraction to model and represent the temporal changes and dynamic aspects of the data underlying many complex systems. Significant insights regarding the stable relational patterns among the entities can be gained by analyzing temporal evolution of the complex entity relations..[2]Web crawlers are essential to many Web applications, such as Web search engines, Web archives, and Web directories, which maintain Web pages in their local repositories. We propose a set of crawling algorithms for effective and efficient crawl ordering by prioritizing important pages with the well-known Page Rank as the importance metric reflected in the gradually descending curves in the performance of semantic focused crawlers. [3] a nature-inspired theory to model collective behavior from the observed data on blogs using swarm intelligence, where the goal is to accurately model and predict the future behavior of a large population after observing their interactions during a training phase. Specifically, an ant colony optimization model is trained with behavioral trend from the blog data and is tested over real-world blogs. Promising results were obtained in trend prediction using ant colony based pheromone classier and CHI statistical measure. [4] Over the past decade, there has been an explosion of interest in network research across the physical and social sciences. For social scientists, the theory of networks has been a gold mine, yielding explanations for social phenomena in a wide variety of disciplines from psychology to economics. Here, we review the kinds of things that social scientists have tried to explain using social network analysis and provide a nutshell description of the basic assumptions, goals, and explanatory mechanisms prevalent in the field. [5] A collective approach to learning a Bayesian network from distributed heterogeneous data. Bayesian network is learnt at the central site using the data transmitted from the local site. The local and central Bayesian networks are combined to obtain a collective Bayesian network, which models the entire data. Experimental results and theoretical justification that demonstrate the feasibility of our approach are presented.

V. PROPOSED WORK

In proposed system to build a stream-based Big Data analytic framework for fast response and real-time decision making.The key challenges and research issues include designing Big Data sampling mechanisms to reduce Big Data volumes to a manageable size for processing; - building prediction models from Big Data streams. Such models can adaptively adjust to the dynamic changing of the data.A knowledge indexing framework to ensure real-time data monitoring and classification for Big Data applications.

VI. PROPOSED ARCHITECTURE/PROTOTYPE



VII. SCOPE OF WORK

Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences. This paper presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution

VIII. DISCUSSION

To explore Big Data, we have analyzed several challenges at the data, model, and system levels. To support Big Data mining, high-performance computing platforms are required, which impose systematic designs to unleash the full power of the Big Data. At the data level, the autonomous information sources and the variety of the data collection environments, often result in data with complicated conditions, such as missing/uncertain values. In other situations, privacy concerns, noise, and errors can be introduced into the data, to produce altered data copies. Developing a safe and sound information sharing protocol is a major challenge. At the model level, the key challenge is to generate global models by combining locally discovered patterns to form a unifying view. This requires carefully designed algorithms to analyze model correlations between distributed sites, and fuse decisions from multiple sources to gain a best model out of the Big Data. At the system level, the essential challenge is that a Big Data mining framework needs to consider complex relationships between samples, models, and data sources,

along with their evolving changes with time and other possible factors. A system needs to be carefully designed so that unstructured data can be linked through their complex relationships to form useful patterns, and the growth of data volumes and item relationships should help form legitimate patterns to predict the trend and future. We regard Big Data as an emerging trend and the need for Big Data mining is arising in all science and engineering domains. With Big Data technologies, we will hopefully be able to provide most relevant and most accurate social sensing feedback to better understand our society at real-time. We can further stimulate the participation of the public audiences in the data production circle for societal and economical events. The era of Big Data has arrived.

IX. REFERENCES

[1] R. Ahmed and G. Karypis, "Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks,"Knowledge and Information Systems,vol. 33, no. 3, pp. 603-630, Dec. 2012.
[2] M.H. Alam, J.W. Ha, and S.K. Lee, "Novel Approaches to Crawling Important Pages Early,"Knowledge and Information Systems,vol. 33, no. 3, pp 707-734, Dec. 2012.
[3] S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks,"Science,vol. 337, pp. 337-341, 2012.
[4] A. Machanavajjhala and J.P. Reiter, "Big Privacy: Protecting Confidentiality in Big Data,"ACM Crossroads,vol. 19, no. 1, pp. 20-23, 2012.
[5] S. Banerjee and N. Agarwal, "Analyzing Collective Behavior from Blogs Using Swarm Intelligence," Knowledge and Information Systems,vol. 33, no. 3, pp. 523-547, Dec. 2012.
[6] E. Birney, "The Making of ENCODE: Lessons for Big-Data Projects,"Nature,vol. 489, pp. 49-51, 2012.
[7] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market,"J. Computational Science,vol. 2, no. 1, pp. 1-8, 2011.
[8] S. Borgatti, A. Mehra, D. Brass, and G. Labianca, "Network Analysis in the Social Sciences,"Science,vol. 323, pp. 892-895, 2009.
[9] J. Bughin, M. Chui, and J. Manyika, Clouds, Big Data, and

Savita Suryavanshi, Bachelor of Engineering in Computer Science and Engineering from Institution of Engineering's Kolkata Currently pursuing Master of Engineering from Pune University.

Prof. Bharati kale, Bachelor of Engineering in Computer Science and Engineering and M.Tech from VTU University, Karnataka.